Unsupervised Learning of Monocular Depth and Ego-Motion using Conditional *Patch*GANs

Madhu Vankadari , **Swagat Kumar** , **Anima Majumder** and **Kaushik Das**
TCS Research and Innovation, Bangalore, India
{madhu.vankadari, swakat.kumar, anima.majumder and kaushik.da}@tcs.com

Abstract

This paper presents a new GAN-based deep learning framework for estimating absolute scale aware depth and ego motion from monocular images using a completely unsupervised mode of learning. The proposed architecture uses two separate generators to learn the distribution of depth and pose data for a given input image sequence. The depth and pose data, thus generated, are then evaluated by a patch-based discriminator using the reconstructed image and its corresponding actual image. The patch-based GAN (or *Patch*GAN) is shown to detect high frequency local structural defects in the reconstructed image, thereby improving the accuracy of overall depth and pose estimation. Unlike conventional GANs, the proposed architecture uses a conditioned version of input and output of the generator for training the whole network. The resulting framework is shown to outperform all existing deep networks in this field, beating the current state-of-the-art method by 8.7% in absolute error and 5.2% in RMSE metric. To the best of our knowledge, this is first deep network based model to estimate both depth and pose simultaneously using a conditional patch-based GAN paradigm. The efficacy of the proposed approach is demonstrated through rigorous ablation studies and exhaustive performance comparison on the popular KITTI outdoor driving dataset.

1 Introduction

Depth and Ego motion estimation from images is an important problem in computer vision which finds application in several fields such as augmented reality [Marchand *et al.*, 2016], 3D re-construction [Geiger *et al.*, 2011], self-driving cars [Handa *et al.*, 2014] etc. Recent advances in deep learning have helped in achieving new benchmarks in this field which is getting better and better with time. The initial deep models [Eigen *et al.*, 2014], [Liu *et al.*, 2016] used supervised mode of learning that required explicit availability of ground truth depth which is not always possible in real world applications. This is partially remedied by using semi-supervised methods which either use sparse ground truth obtained from sensors like LIDAR [Kuznietsov *et al.*, 2017] or make use of synthetically generated data as ground truth [Luo *et al.*, 2018]. Compared to these methods, the unsupervised methods are becoming more popular with time as no explicit ground truth information is required for the learning process. In these cases, the geometric constraints between a pair of images either in temporal [Zhou *et al.*, 2017], [Mahjourian *et al.*, 2018] or spatial domain [Godard *et al.*, 2017] or both [Babu *et al.*, 2018] are exploited to estimate the depth and pose information. Some of the most recent and best results in this category are reported by methods such as, Vid2Depth [Mahjourian *et al.*, 2018], UnDeepVO [Li *et al.*, 2018], Depth-VO-Feat [Zhan *et al.*, 2018] and UnDEMoN [Babu *et al.*, 2018]. Vid2Depth [Mahjourian *et al.*, 2018] uses inferred 3D world geometry and enforces consistency of estimated point clouds and pose information across consecutive frames. Since they rely on temporal consistency (monocular sequence of images), the absolute scale information is lost. This is remedied in UnDeepVO [Li *et al.*, 2018] where authors enforce both spatial and temporal consistencies between images as well as between 3D point clouds. UnDEMoN [Babu *et al.*, 2018] further improves the performance of UnDeepVO [Li *et al.*, 2018] by predicting disparity instead of depth and using a different penalty function for training. Depth-VO-Feat [Zhan *et al.*, 2018] attempts to further

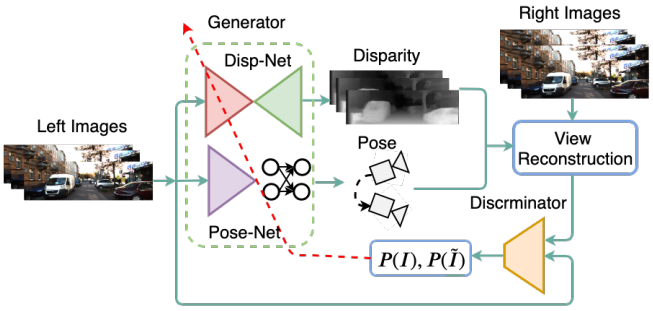

Figure 1: Architectural overview of proposed method. Disp-Net and Pose-Net are used as generators to learn the distribution of depth and pose data respectively for a given input image sequence. Discriminator network evaluates the generator by using the reconstructed images obtained from the view reconstruction module and the original images

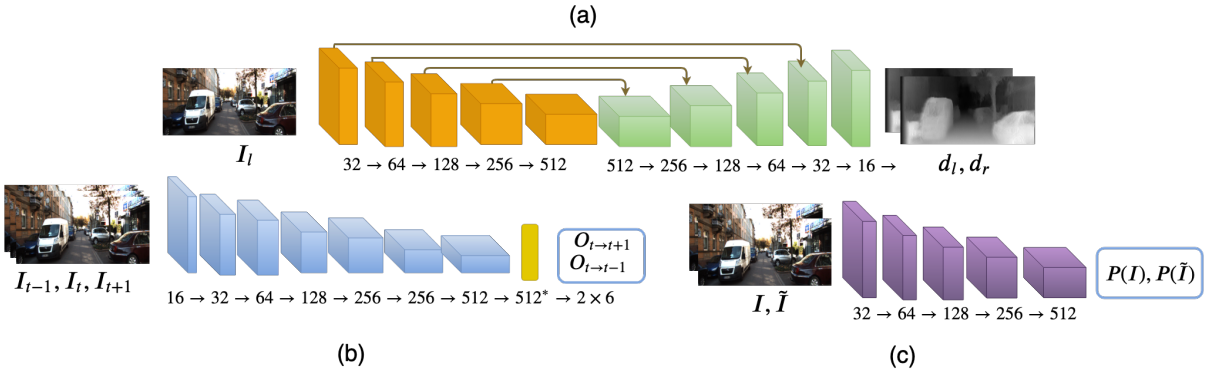

Figure 2: Detailed network architecture for each of the generator and discriminator modules. (a) Disp-Net (b) Pose-net and (c) Discriminator

improve the results by including deep feature-based warping losses into the training process. These deep features are obtained from a depth model that is pre-trained on a different dataset through a supervised mode of learning.

In spite of these advancements, the accuracy of these methods are still far from those obtained using stereo methods [Godard *et al.*, 2017] or supervised methods [Luo *et al.*, 2018] leaving enough room for further improvements. With this hindsight, we restrict our discussion only to unsupervised methods in this paper with an aim to produce superior performance creating new benchmark in this field.

Most of the unsupervised methods make use of image reconstruction losses computed in spatial or temporal domains to learn the mapping from pixel to depth and pose information. The image reconstruction losses are usually obtained by comparing the reconstructed images with their corresponding original images using metrics such as $L_{p;(p=1,2)}$ norm [Zhou *et al.*, 2017], $SSIM$ [Godard *et al.*, 2017] etc. Rather than directly computing these image reconstruction losses, one can also use a discriminative network to directly evaluate if the reconstructed image is good or bad. Such discriminative networks are an integral part of Generative Adversarial Networks (GAN) [Goodfellow *et al.*, 2014] which are commonly used in solving problems like image-to-image translation [Isola *et al.*, 2017], text-to-image synthesis [Reed *et al.*, 2016], style transfer [Zhu *et al.*, 2017] etc. A generative adversarial network (GAN) consists of two networks - a generator model that tries to mimic the underlying distribution of input data and, a discriminator network that learns to evaluate this distribution given the original distribution. These two networks help each other improve their performance by playing a zero-sum game. The advantage of such a paradigm is that one can generate faithful synthetic data necessary for learning an input-output map in cases where the actual real-world data is unavailable or scarce. This has prompted many researchers to make use of GANs for improving the accuracy of depth and pose estimation from monocular as well as stereo images [Kumar *et al.*, 2018], [Aleotti *et al.*, 2018], [Pilzer *et al.*, 2018], [Almalioglu *et al.*, 2019]. In these works, the depth and pose estimation network is used as a generator which is expected to produce accurate depth or pose information. The output of the generator is then evaluated by a discriminator that uses the actual image and the reconstructed image obtained from the above estimated depth and pose information. In other words, a conditioned version of the generator output is used by the discriminator for evaluation. Again, the generators used in these cases are a conditional one as they use a pre-defined set of images to produce the depth or pose data instead of a generic random noise signal.

In this paper, we adopt the above conditional GAN paradigm for estimating absolute scale aware depth and pose information from monocular stereo images. Unlike the above methods that use a scalar value to decide whether the output of the generator is real or fake (good or bad), we propose to use a patch-based discriminator that evaluates an array of smaller patches of the reconstructed image instead of the whole image. This is otherwise known as *Patch*GAN [Isola *et al.*, 2017] which is shown to capture high frequency local structural information in the reconstructed image which, otherwise, gets ignored when a $L1/L2$ kind of loss function is used. In addition, the *Patch*GAN is fully convolutional in nature which makes it lighter and faster compared to others who use dense or fully connected layers in their architecture. Based on this intuition, we have carefully designed a new GAN-based deep network architecture that includes a generator for depth and pose information and a patch-based discriminator for evaluating the generator performance in an adversarial mode of learning. Some of the other features of this architecture are as follows. The generator input is conditioned over a given image to produce the depth or pose information unlike methods which design prior distributions to this effect [Almalioglu *et al.*, 2019] or use a random noise as done in a traditional GAN approach. Secondly, the proposed architecture uses a single discriminator to evaluate both depth and pose generators making it novel and unique in this field. The resulting effect of these features can be appreciated from the fact that the proposed model outperforms all existing state-of-the-art methods to create a new benchmark in the field. The efficacy of the proposed approach is demonstrated through extensive analysis and ablation studies carried out using the popular KITTI driving dataset [Geiger *et al.*, 2013].

In short, our major contribution lies in proposing a new conditional *Patch*GAN-based deep network architecture for estimating absolute scale aware depth and pose information from monocular images which is shown to provide state-of-

the-art performance in this field while being one of the lightest and fastest GAN model, in terms of trainable parameters, reported so far in the literature.

The rest of the paper is organized as follows. The proposed method along with the network architectures is explained in the next section. The details of experiment and analysis of results are presented in Section 3. The conclusion and direction for future work is provided in Section 4.

2 Proposed Method

The GANs are generative models with an ability to learn mapping between a random noise vector z to an output image y, $G : z \rightarrow y$ [Goodfellow *et al.*, 2014]. In contrast, the conditional GANs learn a mapping from an user-defined image x and random noise vector z to an output image y, $G : \{x, z\} \rightarrow y$ [Mirza and Osindero, 2014]. The conditional GAN has the advantage of directing generator's output towards a particular context by taking additional information as input and hence, can be used to deal with applications that require one-to-many mapping (e.g., a single image can be tagged with multiple labels by different human beings). We make use of this understanding to generate depth (disparity) and pose data conditioned to a particular input image sequence, rather than generating from a random noise vector. The proposed deep network architecture for estimating absolute scale aware depth and pose information from a monocular sequence of images based on this concept of conditional GAN is shown in Figure 1. The generator module G consists of two deep networks - one for estimating the disparity and the other for estimating the pose or the ego-motion. The disparity and the pose information, thus obtained, are then used by the view reconstruction module V to reconstruct the target images. These reconstructed images are evaluated by a fully convolutional patch-based discriminator D [Isola *et al.*, 2017] by comparing them with the corresponding original images. Each of these modules are described next in this section below.

2.1 Generator Module: Disparity and Pose Estimation Networks

Estimating depth directly from images has two major disadvantages. First, the uncertainty in depth prediction increases with increasing distance in the scene. Secondly, the value of depth at horizon goes to infinity making it difficult to compute the gradient values needed for training a deep network. These two problems are remedied by designing the deep network to predict disparity (inverse of depth) instead of predicting depth directly [Godard *et al.*, 2017], [Zhou *et al.*, 2017], [Babu *et al.*, 2018], [Zhan *et al.*, 2018]. Keeping in mind the real world implementation requirement, we use a trimmed version of the original disparity network with only five convolutional layers, instead of the standard seven layer network used by the previous researchers [Godard *et al.*, 2017], [Zhou *et al.*, 2017]. It has about 8 million trainable parameters which is only one-fourth of the original network size, making it one of the lightest and thinnest deep network model for depth estimation. The details of the modified disparity network DispNet is shown in Figure 2(a). Given a dataset of stereo images $X = \{I_l, I_r\}$, the Disp-Net takes the left image I_l as input to predict the left-to-right disparity d_l and right-to-left disparity d_r. Once the disparity is known, the depth can be calculated by using the intrinsic and extrinsic calibration parameters of the stereo-rig as $\hat{d} = bf/d$, where f is the focal length, b is the baseline distance between the stereo camera pair and $d\,\epsilon\{d_l,\ d_r\}$ is the predicted disparity.

The Pose-Net is a convolutional encoder followed by a fully connected layer as shown in Figure 2(b). This network takes a snippet of n-temporally aligned monocular images as input and predicts the relative ego-motion O which includes translation (t_x, t_y, t_z) and rotation (ρ, θ, ψ) of the camera between the frames of the snippet.

These two deep networks form the generator module for the proposed architecture which aims to learn the distribution of disparity and pose data for a given input image sequence. The authenticity of these generator outputs are evaluated by a discriminator module, details of which will be discussed later in this section.

2.2 View Reconstruction Module

The view reconstruction module V has two sub-modules namely a spatial reconstruction module S and a temporal reconstruction module T. The spatial reconstruction S module takes the predicted disparities $\{d_l, d_r\}$ and reconstructs the left and right images $\tilde{I}_l, \tilde{I}_r$ using the original pair $\{I_r, I_l\}\epsilon X$ through inverse-image warping technique. The temporal reconstruction module T takes a sequence of temporally-aligned consecutive image frames (n-snippet), the estimated depth of the center frame and the predicted pose between these frames to reconstruct the target image. For $n = 3$, the view reconstruction module V takes an input snippet (I_{t-1}, I_t, I_{t+1}) along with estimated depth of center frame $\hat{d}_t$ and the predicted pose $(O_{t\rightarrow t+1}, O_{t\rightarrow t-1})$ to reconstruct the target image $\tilde{I}_t$ from the source images I_{t+1}, I_{t-1}. The reconstruction modules $V = (S, T)$ uses bi-linear interpolation of [Jaderberg *et al.*, 2015] to add RGB color intensities to the transformed pixels of the reconstructed image. Please refer to [Babu *et al.*, 2018] for more information regarding the reconstruction modules.

2.3 *Patch*GAN-based Discriminator

In a traditional GAN paradigm, the discriminator network is used to evaluate the generator output for their authenticity (real or fake) based on the available ground truth. This is not possible in our case where the generator gives out disparity and pose information as output for which no ground truth information is available. Instead, the discriminator is made to evaluate the reconstructed images $(\tilde{I}_l, \tilde{I}_r, \tilde{I}_t)$ obtained from the view reconstruction module V. In other words, the discriminator is using a conditioned version of the generator output for evaluation. The proposed discriminator module makes use of the concept of *Patch*GAN originally proposed in [Isola *et al.*, 2017] for image-to-image translation. Unlike a regular GAN discriminator which produces a scalar value for the entire image signifying the image being real or fake, the *Patch*GAN based discriminator outputs $m \times n$ array of scalars where each signifies the authenticity of a region or

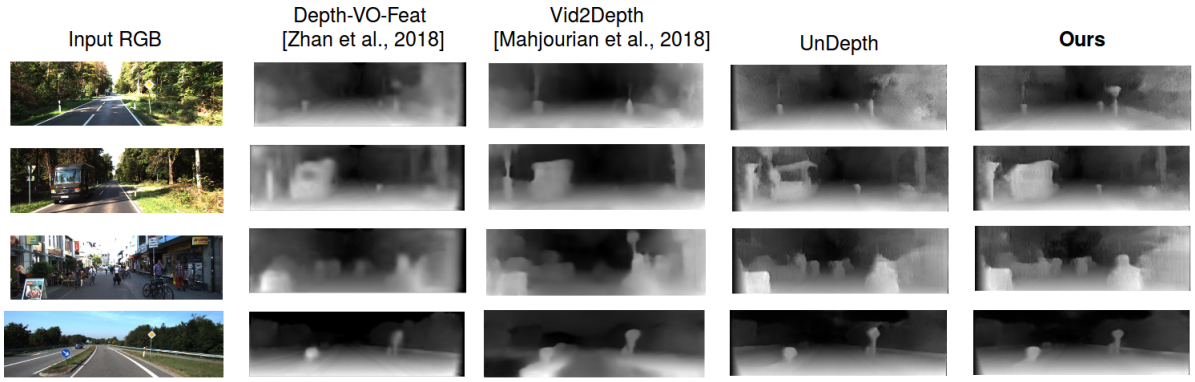

Figure 3: Qualitative comparison of the proposed method with other existing state-of-the-art methods. 'Ours' is the proposed *Patch*GAN-based depth and pose model. 'UnDepth' represents the proposed Disp-Net without adversarial training. As one can see, Our method provides sharper images with more details compared to others

patch of the input image. The *Patch*GAN based discriminator is shown to capture the high frequency structural information of local patches allowing it to effectively evaluate generator output. The patch sizes can be made smaller providing high level of discrimination at the cost of increased computational cost. Moreover, this discriminator consists of only convolutional layers making it faster and lighter in terms of trainable parameters. The details of the discriminator network is shown in Figure 2(c). It takes the image pair $(\tilde{I}, I)$ and produces the probability pair $(P(\tilde{I}), P(I))$ which indicates whether the reconstructed image is real or fake.

2.4 Training Losses

The proposed method is an unsupervised approach which does not require explicit ground truth depth and pose information to train the network. Instead, the image reconstruction loss is commonly used for training the network [Godard *et al.*, 2017], [Zhou *et al.*, 2017], [Babu *et al.*, 2018]. The image reconstruction loss used for the proposed GAN-based architecture has two components, namely, (1) *content loss* that enforces geometry (appearance) and regularization (smoothness and left-right consistency) into the network, and (2) the *adversarial loss* obtained from the discriminator that returns a probability of a given reconstructed image being real or fake. The details of these two losses are described next in this section.

Content Losses

The content losses include appearance loss, smoothness loss and consistency loss which are defined as follows. The appearance loss is calculated both in spatial and temporal domains [Babu *et al.*, 2018]. The spatial appearance loss is computed by comparing the reconstructed left-right image pairs $(\tilde{I}_l, \tilde{I}_r)$ with their corresponding original images (I_l, I_r). Similarly, the temporal loss is computed by comparing the reconstructed target image $\tilde{I}_t$ with the original target image I_t. For example, the image appearance loss for a given original left image I_l and the corresponding reconstructed left image $\tilde{I}_l$ is given by

$$L^l_{ap} = \frac{1}{N}\sum_{ij} \alpha\rho[\frac{1 - SSIM(I^{ij}_l, \tilde{I}^{ij}_l)}{2}] + (1-\alpha)\rho(||I^{ij}_l - \tilde{I}^{ij}_l||) \quad (1)$$

where $\rho(.)$ is the Charbonnier Penalty function [Babu *et al.*, 2018] and $SSIM(.)$ is the structural similarity index between the original image and the reconstructed image [Godard *et al.*, 2017]. The parameter $\alpha < 1$ is the relative weight given to each of the components. The symbol I^{ij}_l represents the intensity value of each pixel (i, j) of the image I_l.

The smoothness loss is calculated by using the predicted disparity and its respective input image for the network. The smoothness loss enforces the predicted disparities to be locally smooth and this is achieved by weighing the disparity gradients (∂d) with exponentially weighted image gradients (∂I). Mathematically, the smoothness loss is given by

$$L^l_{ds} = \frac{1}{N}\sum_{ij} \rho(\partial_x d^{ij}_l e^{-||\partial_x I^{ij}_l||}) + \rho(\partial_y d^{ij}_l e^{-||\partial_y I^{ij}_l||}) \quad (2)$$

The consistency loss [Godard *et al.*, 2017] enforces cycle consistency between the predicted disparities by projecting disparities from one to the other. Mathematically, this is given by

$$L^l_{lr} = \frac{1}{N}\sum_{ij} \left\| d^{ij}_l - d^{ij+d^{ij}_l}_r \right\| \quad (3)$$

2.5 Adversarial Loss

As mentioned earlier, the discriminator module D takes the reconstructed images obtained from the view reconstruction module and their corresponding real images to evaluate the disparity and pose output of the generator module. The discriminator assigns probability $P(.)$ to each image a value that ranges from 0 to 1, 0 being completely fake and 1 being completely real. The objective of the generator G is to fool the discriminator D while the discriminator tries not to get fooled by the generator by correctly labeling the images. This is achieved by the following objective function which

represents the min-max game played between the generator and the discriminator [Goodfellow *et al.*, 2014]:

$$\begin{aligned}\mathcal{L}_{GAN}(G,D) &= \min_G \max_D U(G,D) \quad (4)\\ &= \mathbb{E}_{i_o \epsilon I}[logD(i_o)] + \mathbb{E}_{i_r \epsilon \tilde{I}}[log(1-D(i_r))]\end{aligned}$$

where i_o and i_r belong to real (I) and reconstructed images ($\tilde{I}$) respectively. In our case, the generator and the discriminator are trained to minimize the following loss functions:

$$L_G = \beta_g \mathbb{E}_{i \epsilon \tilde{I}}[logD(i_r)] + L_{content} \quad (5)$$

$$L_D = \mathbb{E}_{i_o \epsilon I}[logD(i_o)] + \mathbb{E}_{i_r \epsilon \tilde{I}}[log(1-D(i_r))] \quad (6)$$

where $L_{content}$ is the content loss given by

$$L_{content} = \beta_{ap} L_{ap} + \beta_{ds} L_{ds} + \beta_{lr} L_{lr} \quad (7)$$

where all βs are the weights given to the individual components emphasizing the trade-off among them.

3 Experiments and Results

The proposed method is implemented in Tensorflow architecture. The total number of trainable parameters in this model is around 12 million and the network is trained for 240k iterations on Dell Alienware laptop having NVIDIA GTX 1080 GPU with 8 GB of GPU memory. Leaky-ReLu [Xu *et al.*, 2015] activation functions are used for all the network layers and the network is optimized by using Adam [Kingma and Ba, 2014] as the optimization algorithm. The learning rate for training is initially set to 0.0001, then it is reduced by half after 3/5th of the iterations and further reduced by half after 4/5th of iterations. The γ value of the Charbonnier penalty [Babu *et al.*, 2018] is set to 0.45. The α value in appearance loss is set to 0.85. The appearance loss and left-right consistency loss weights β_{ap} and β_{lr} are set to 1.0. The smoothness loss weight β_{ds} is set to $0.1/s$ where s is the ratio of respective disparity image resolution to the input image resolution. The adversarial loss β_g is set 0.001 which is obtained after extensive ablation study.

3.1 KITTI Dataset

The KITTI dataset is a popular outdoor driving dataset containing 61 different driving sequences with 42382 images of resolution 1242×345. The dataset is divided into two splits namely, KITTI-Stereo split and KITTI-Eigen split which are commonly used to benchmark the performance of algorithms for depth and pose estimation. More details about the dataset and its use can be found in [Babu *et al.*, 2018], [Godard *et al.*, 2017] which are being omitted here for the sake of brevity. The performance analysis of the proposed method on this dataset is described next.

3.2 Depth Evaluation

The depth evaluation is performed on both stereo and Eigen splits. We have used standard evaluation metrics of [Eigen *et al.*, 2014] for comparison with other existing methods. The qualitative comparison of various methods is shown in Figure 3. As one can see, our proposed model provides sharper images with more details compared to other methods. This is further confirmed through the quantitative comparison analysis presented in Table 1 where the performance results are reported for the KITTI eigen and stereo splits. The first two parts of the tables show the results of Eigen split with maximum depth range of 80m and 50m and the lower part shows the results on Stereo-split with 80m as the maximum depth range. The results of other methods are directly taken from their respective papers. The 'UnDepth' is the proposed Disp-Net model which is trained with unsupervised mode of learning only using the content loss $L_{content}$. The method 'Undepth+PatchGAN' refers to case where only Disp-Net is trained in adversarial fashion while the method 'Undepth+Pose+PatchGAN' is our proposed method that trains both Disp-Net and Pose-Net using adversarial learning. As one can observe, the proposed method outperforms all existing methods in both error and accuracy metrics.

3.3 Ablation Study

We have performed two ablation studies. The first one is done for finding the suitable value of the parameter β_g used for weighing the adversarial loss during training with a discriminator that has 5 convolutional layers. This parameter is observed to be very sensitive and greatly affects the performance of the overall network. The resulting outcome is shown quantitatively in Table 2. Based on this study, the value of β_g is selected to be 0.001 in order to report final performance measures for comparison in Table 1. The second ablation study is carried out to decide the number of layers to be used for the discriminator network which in turn effects the size of the patch being used for evaluating the generator output. The resulting effect of this parameter is shown in second part of the Table 2. This study shows that the network provides best performance with five convolutional layers.

3.4 Pose Evaluation

The performance of Pose-Net is evaluated using the image sequences of the Odometry split which are in the test set of the eigen split of the KITTI dataset as explained in [Babu *et al.*, 2018]. We use Absolute Trajectory Error [Zhou *et al.*, 2017], [Babu *et al.*, 2018] as a measure for comparing the performance of our model with other state-of-the-art methods in the field. The resulting comparison is shown in Table 3. The SfMLearner [Zhou *et al.*, 2017] employs a post processing stage that uses ground truth pose to obtain the absolute scale information and is referred to by using the suffix _PP. For a fair comparison with our method that does not use any ground truth, we obtain the results for SfMLearner by removing this post processing step and is denoted by the suffix _noPP. Similarly, we compare the performance of our algorithm with the monocular (VISO_M) and stereo (VISO_S) version of the VISO [Geiger *et al.*, 2011] model which is a known traditional method in this category. As one can see the proposed method outperforms UnDEMoN, SfMLearner_noPP and VISO_M and is comparable to the VISO_S and SfMLearner_PP that use ground truth information explicitly.

3.5 Discussion

1. The proposed network architecture uses both left and right images during the training phase to predict dis-

Method	Supervision	Abs Rel	Sq Rel	RMSE	logRMSE	δ <1.25	δ <1.25^2	δ <1.25^3
Train set mean	D	0.361	4.826	8.102	0.377	0.638	0.804	0.894
SfM Learner** [Zhou *et al.*, 2017]	M	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Monodepth [Godard *et al.*, 2017]	MS	0.148	1.344	5.927	0.247	0.803	0.922	0.964
UnDeepVO [Li *et al.*, 2018]	MS	0.183	1.73	6.57	0.283	-	-	-
Vid2Depth** [Mahjourian *et al.*, 2018]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GANVO** [Almalioglu *et al.*, 2019]	M	0.150	1.141	5.448	**0.216**	0.808	**0.939**	**0.975**
Depth-VO-Feat [Zhan *et al.*, 2018]	MS	0.144	1.391	5.869	0.241	0.803	0.928	0.969
[Kumar *et al.*, 2018]	M	0.2114	1.9797	6.1540	0.263	0.731	0.897	0.959
[Pilzer *et al.*, 2018]	S	0.152	1.388	6.016	0.247	0.789	0.918	0.965
UnDEMoN [Babu *et al.*, 2018]	MS	0.139	1.174	5.59	0.239	0.812	0.930	0.968
UnDepth	MS	0.1365	1.1391	5.642	0.239	0.813	0.928	0.967
UnDepth + PatchGAN	MS	0.1306	1.076	5.470	0.231	0.821	0.933	0.971
UnDepth+Pose+PatchGAN	MS	**0.1269**	**0.9982**	**5.309**	**0.226**	**0.827**	**0.934**	**0.971**
Monodepth [Godard *et al.*, 2017]	MS	0.140	0.976	4.471	0.232	0.818	0.931	0.969
SfM Learner** [Zhou *et al.*, 2017]	M	0.201	1.391	5.181	0.264	0.696	0.900	0.966
Vid2Depth** [Mahjourian *et al.*, 2018]	M	0.155	0.927	4.549	0.231	0.781	0.931	0.975
Depth-VO-Feat [Zhan *et al.*, 2018]	MS	0.135	0.905	4.366	0.225	0.818	0.937	0.973
UnDEMoN [Babu *et al.*, 2018]	MS	0.132	0.884	4.290	0.226	0.827	0.937	0.972
UnDepth	MS	0.129	0.8344	4.259	0.225	0.827	0.935	0.972
UnDepth+PatchGAN	MS	0.1239	0.7908	4.162	0.218	0.835	0.940	0.974
UnDepth+Pose+PatchGAN	MS	**0.1207**	**0.7490**	**4.051**	**0.214**	**0.840**	**0.941**	**0.975**
Monodepth [Godard *et al.*, 2017]	MS	0.124	1.388	6.125	0.217	0.841	0.936	0.975
MonoGAN [Aleotti *et al.*, 2018]	MS	0.119	1.239	5.998	0.212	0.846	0.940	0.976
UnDepth	MS	0.1192	1.2891	5.959	0.214	0.840	0.937	0.974
UnDepth+PatchGAN	MS	0.1161	1.317	5.781	0.206	**0.850**	0.944	0.978
UnDepth+Pose+PatchGAN	MS	**0.1102**	**1.0443**	**5.535**	**0.200**	0.849	**0.944**	**0.979**

Table 1: Performance Comparison of the proposed method with the other state-of-the-art techniques using Eigen (80m, 50m) and stereo splits. The first two parts of the table shows the results for Eigen split with 80m and 50m maximum depth range respectively. The bottom part of the table shows results for the stereo split with 80m maximum depth range. The column Supervision indicates the type of supervision used for training, where D refers to Depth, M to Monocular and MS refers to Monocular Stereo

β_g	Abs Rel	Sq Rel	RMSE	logRMSE	δ <1.25	δ <1.25^2	δ <1.25^3
0.1	0.1327	1.088	5.497	0.232	0.821	0.932	0.970
0.01	0.1337	1.100	5.511	0.232	0.819	0.931	0.970
0.001	**0.1306**	**1.076**	**5.470**	**0.231**	**0.821**	**0.933**	**0.971**
0.0001	0.1338	1.083	5.479	0.233	0.817	0.932	0.970
# Conv layers							
upto Conv4	0.1354	1.0942	5.477	0.233	0.820	0.932	0.970
upto **Conv5**	**0.1306**	**1.076**	**5.470**	**0.231**	**0.821**	**0.933**	**0.971**
upto Conv6	0.1311	1.0998	5.494	0.231	0.823	0.932	0.970

Table 2: The Ablation study for selecting the value of β_g and the number of convolutional layers used in the Discriminator model

Seq	**Ours**		UnDEMoN		SfMLearner_noPP		SfMLearner_PP		VISO2_S		VISO_M	
	t_{ate}	r_{ate}	t_{ate}	r_{ate}	t_{ate}	r_{ate}	t_{ate}	r_{ate}	t_{ate}	r_{ate}	t_{ate}	r_{ate}
00	0.0593	0.0013	0.0644	0.0013	0.7366	0.0040	0.0479	0.0044	0.0429	0.0006	0.1747	0.0009
04	0.0713	0.0006	0.0974	0.0008	1.5521	0.0027	0.0913	0.0027	0.0949	0.0010	0.2184	0.0009
05	0.0651	0.0008	0.0696	0.0009	0.7260	0.0036	0.0392	0.0036	0.0470	0.0004	0.3787	0.0013
07	0.0666	0.0010	0.0742	0.0011	0.5255	0.0036	0.0345	0.0036	0.0393	0.0004	0.4803	0.0018

Table 3: Absolute Trajectory Error (ATE) for Translation and Rotation on KITTI eigen split dataset averaged over all 3-frame snippets (lower is better). As one can see, our method outperforms the monocular versions SfMLearner_noPP and VISO_M and is comparable with stereo versions SfMLearner_PP and VISO_S . Here, the terms t_{ate} and r_{ate} stand for translational absolute trajectory error and rotational absolute trajectory error respectively

parity which is then used for computing depth by using the extrinsic and intrinsic calibration parameters of the stereo-camera rig used for acquiring these images. During the testing phase, the network only takes left image as an input to produce scale aware depth as output. Given an entirely new dataset with different camera setup, the trained model is observed to predict depth and pose which is correct upto a scale while retaining all the structural attributes of the scene in the disparity image. The actual scale can be retrieved by re-training (fine-tuning) the network on the new dataset (with minor additional computational cost) for a fewer number of epochs.

2. It should be noted that the methods like SfMLearner

[Zhou *et al.*, 2017], GANVO [Almalioglu *et al.*, 2019] and Vid2Depth [Mahjourian *et al.*, 2018] have used approximately 40000 images for training, unlike our model which is trained on only 22600 images as per the prevailing practice [Babu *et al.*, 2018], [Godard *et al.*, 2017], [Pilzer *et al.*, 2018]. Ideally, it is not fair to directly compare their performance with ours as they also include test images into the training set. The performance parameters for these three algorithms have been marked with double asterisk (**) symbol in Table 1. We have included these results to show that the proposed network produces better result compared to these methods even with a smaller training dataset. Furthermore, this superior performance is obtained using a network size which is approximately one-fourth of the networks used previously [Godard *et al.*, 2017], [Babu *et al.*, 2018], [Zhou *et al.*, 2017].

4 Conclusions

The paper presents a novel deep network architecture based on conditional *Patch*GANs for estimating absolute scale aware depth and pose information from monocular images. The proposed GAN architecture uses disparity and pose estimation network as generators and a fully convolutional network as a discriminator. The discriminator evaluates the local patches of the reconstructed image in order to evaluate the generator output. This has the advantage of capturing local structural information which usually get lost when only one scalar value is used to evaluate the whole image. The proposed deep architecture uses only one-fourth of the number of trainable parameters compared to other deep networks reported in the literature. The resulting framework is shown to out-perform all existing methods to create new benchmark in this field. However, the proposed model is not capable of dealing with moving objects and occlusions. In addition, the long term dependencies in pose estimation for loop closure has not been considered. These problems form the future scope of this work.

References

[Aleotti *et al.*, 2018] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In *15th European Conference on Computer Vision (ECCV) Workshops*, volume 1, page 8, 2018.

[Almalioglu *et al.*, 2019] Yasin Almalioglu, Muhamad Risqi U Saputra, Pedro PB de Gusmao, Andrew Markham, and Niki Trigoni. Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[Babu *et al.*, 2018] V Madhu Babu, Kaushik Das, Anima Majumdar, and Swagat Kumar. Undemon: Unsupervised deep network for depth and ego-motion estimation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1082–1088. IEEE, 2018.

[Eigen *et al.*, 2014] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[Geiger *et al.*, 2011] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium*, Baden-Baden, Germany, June 2011.

[Geiger *et al.*, 2013] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[Godard *et al.*, 2017] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611. IEEE, 2017.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[Handa *et al.*, 2014] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *Robotics and automation (ICRA), 2014 IEEE international conference on*, pages 1524–1531. IEEE, 2014.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[Kumar *et al.*, 2018] Aran CS Kumar, Suchendra M Bhandarkar, and Mukta Prasad. Monocular depth prediction using generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 413–4138. IEEE, 2018.

[Kuznietsov *et al.*, 2017] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6647–6655, 2017.

[Li *et al.*, 2018] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7286–7291. IEEE, 2018.

[Liu *et al.*, 2016] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2016.

[Luo *et al.*, 2018] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018.

[Mahjourian *et al.*, 2018] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.

[Marchand *et al.*, 2016] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2016.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Pilzer *et al.*, 2018] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *2018 International Conference on 3D Vision (3DV)*, pages 587–595. IEEE, 2018.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[Xu *et al.*, 2015] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[Zhan *et al.*, 2018] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018.

[Zhou *et al.*, 2017] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.